APML - Clustering

יובל יעקבי, 302247077

18 בדצמבר 2018

חלק תאורטי

שאלה 1.1 - התכנסות של kmeans

נזכר באלגוריתם - מטרתו לחלק n נקודות לk קלאסטרים, כאשר k, n קבועים
ראשית נשים לב שיש מספר סופי של חלוקות שונות של n נקודות לk קלאסטרים - k^n
כעת נראה שפונקציית המחיר - $\sum_{i=1}^{k}\sum_{x\in C_i}||x-\mu_i||^2$ מונטונית יורדת זה אומר בהכרח שמתישהו נעצור כיוון שיש מספר סופי של מצבים.
יש לנו 2 צעדים לאלגוריתם:

1. בחירת מרכז - בצעד הזה כל אחת מהנקודות משויכת לאחד המרכזים ואנחנו בוחרים את המרכז, כיוון שאנחנו בוחרים את המרכז כפונקציית אופטימיזציה שמשיגה מינימום על $\sum_{x\in C_i}||x-\mu_i||^2$ זה בהכרח יכול רק לרדת
2. שלב הקלאסטרים - בשלב הזה כל נקודה נבחרת לאחד הקלאסטרים, גם כאן כל נקודה משויכת למרכז הקרוב ביותר אליה - כלומר $C(x) = argmin_k ||x-\mu_k||^2$ גם כן זו בעיית אופטימיזציה של מינימום ולכן פונקציית המחיר יכולה רק לרדת...

בסה"כ בכל צעד פונקציית המחיר לא גדלה (כלומר קטנה או נשארת במקום), וכיוון שיש מספר סופי של מצבים מתישהו בהכרח נגיע למצב שנשאר במקום ולכן נעצור...

שאלה 1.2 - kmeans לא אופטימלי

המרחק שנעבוד איתו הוא: $\sum_{i=1}^{k}\sum_{x\in C_i}||x-\mu_i||^2$

1. דוגמא לאתחול לא טוב - נחשוב על מלבן צר וארוך כלומר נקודות הן: $\{(0,0),(0,10),(1,0),(1,10)\}$

נניח שהאתחול היה בדיוק באמצע הצלע של כל אחת מהצלעות הארוכות, כלומר $\left\{\hat{a}=(0,5),\hat{b}=(1,5)\right\}$ (לצורך נוחות סימנו את שמות הקלאסטרים ב$\hat{a},\hat{b}$)

בשלב הבא נקבל את ההשמות הבאות:

$$\{(0,0),(0,10)\}\in C_{\hat{a}} \wedge \{(1,0),(1,10)\}\in C_{\hat{b}}$$

בשלב הבא מרכזי הקלאסטרים לא יזוזו כיוון שהמרכז הוא בדיוק באמצע של הנקודות.

ההשמה הזו כל נקודה נמצאת במרחק של $(5)^2$ ממרכז הקלאסטר שלה, כלומר סה"כ המרחק הוא 100

אם נבחר מרכזי קלאסטרים: $\{a=(0.5,0)\,,b=(0.5,10)\}$, הקלאסטרים שנקבל יהיו אופטימליים עם מרחק $\left(\frac{1}{2}\right)^2$ מהמרכז (לכל נקודה), כלומר סה"כ המרחק הוא 1 ההשמה:

$$\{(0,0)\,,(1,0)\}\in C_a \wedge \{(0,10)\,,(1,10)\}\in C_b$$

2. דוגמא עם יותר ממרכז אופטימלי אחד: נתבונן בריבוע $\{(0,0)\,,(0,1)\,,(1,0)\,,(1,1)\}$ יש לנו 2 קבוצות של מרכזים שישיגו מרחק $\left(\frac{1}{2}\right)^2$ לכל נקודה - סה"כ 1 השמות:

   (א) $\{(0,0.5)\,,(1,0.5)\}$ - כלומר הנקודות שחולקות את אותו ציר y יהיו ביחד

   (ב) $\{(0.5,0)\,,(0.5,1)\}$ - כלומר הנקודות שחולקות את אותו ציר x יהיו ביחד

שאלה 1.3 - איתחול של kmeans

נגדיר - "מאורע טוב" - מאורע שבו האיתחול (הרנדומי) של kmenas בחר כל מרכז בקלאסטר שונה.

נסמן ב$\alpha_1,\dots,\alpha_k$ את הסיכוי לדגום נקודה מכל אחד מk הקלאסטרים

1. נניח שאנחנו דוגמים באקראי נקודות לפי הסתברות אחידה, אזי מאורע טוב הוא:

$$k!\Pi_{i=1}^{k}\alpha_i$$

נסביר - הסדר לא משנה, לכן אנחנו כופלים ב$k!$ (עושים k בחירות, בכל שלב לא אכפת לנו איזה אחד מהקלאסטרים נבחר)

כפל - אנחנו רוצים לבחור אחד מכל קלאסטר, כלומר אנחנו רוצים לבחור: $\alpha_1 \wedge \alpha_2 \wedge \dots \wedge \alpha_k$, כלומר בהסתברות - כפל.

כעת אנחנו מתעניינים בתוחלת, כאמור זהו משתנה מקרי גאומרטי ולכן:

$$\mathbb{E}\left[X\right]=\frac{1}{k!\Pi_{i=1}^{k}\alpha_i}$$

כאשר X הוא מספר הניסיונות שצריך לעשות עד שנקבל הגרלה מוצלחת

2. הסיכוי הכי טוב שלנו הוא כאשר כל הקלאסטרים יהיו עם אותה הסתברות, כלומר $\alpha_1=\alpha_2=\dots=\alpha_k=\frac{1}{k}$

הסיבה לכך היא שאנחנו רוצים לבחור אחד לכל קלאסטר, אם יהיה קלאסטר שהסיכוי שלו גדול יותר, אזי זה אומר בהכרח שיהיה אחד שהסיכוי שלו קטן יותר ויש פחות סיכוי שנבחר בו.

כלומר

$$\mathbb{E}\left[X\right]\geq\frac{1}{k!\Pi_{i=1}^{k}\frac{1}{k}}=\frac{k^k}{k!}\overset{stirling}{\approx}\frac{k^k}{\frac{k^k}{e^k}}=e^k$$

3. זוהי כמובן לא התוצאה שהיינו רוצים לקבל, כלומר אם אנחנו רוצים 10 קלאסטרים נצטרך להריץ את האלגוריתם $e^{10} \approx 22,000$ פעמיים ולבחור את התוצאה הטובה ביותר (וכמובן שזה גדל אקספוננציאלית).

   כמו כן המצב בחיים האמיתים כנראה גרוע יותר כי אני לא חושב שההתפלגות בין קלאסטרים תיהיה אחידה.

   לעומת זאת, החישוב הזה נעשה אם אנחנו בוחרים נקודה מכל קלאסטר, אבל אין סיבה שהמרכז של קלאסטר לא "ינדוד" במהלך האיטרציות של האלגוריתם, כלומר גם אם התחלנו עם 2 מרכזים באותו קלאסטר זה לא אומר שלא נקבל תוצאה אופטימלית, בנוסף לכך הרבה פעמיים נסתפק במקסימום לוקאלי ולא גלובלי, כלומר אנחנו יכולים לקבל קלאסטרים ממש טובים ולא אופטימלים כי פספסנו כמה נקודות וזה עדיין בסדר.

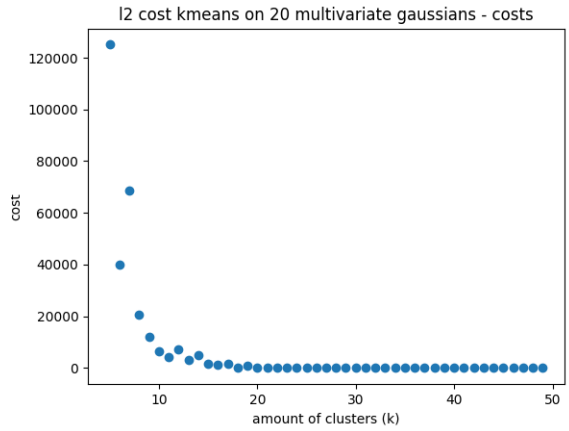
   וכמובן שיש לנו kmeans++ (וכנראה אלגוריתמים אחרים שלא למדנו :))

חלק פרקטי

לאורך התרגיל רשמתי kmeans , הכוונה כמובן לkmeans++

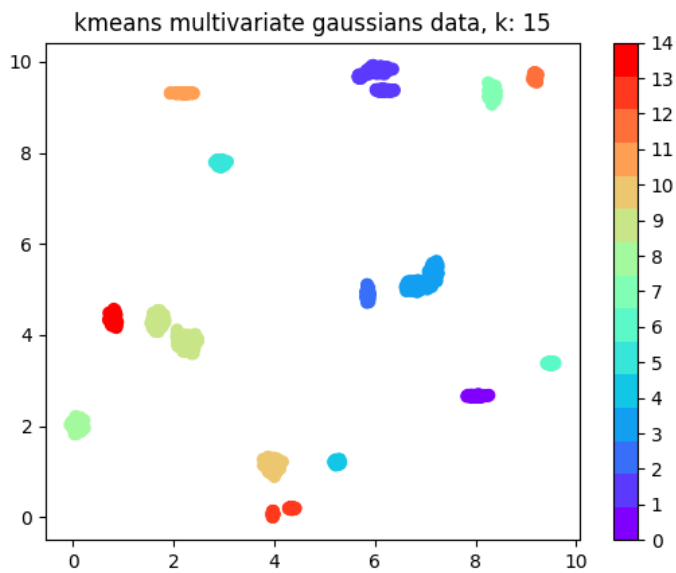
בחינת K־שונים ־ אלגוריתם Kmeans

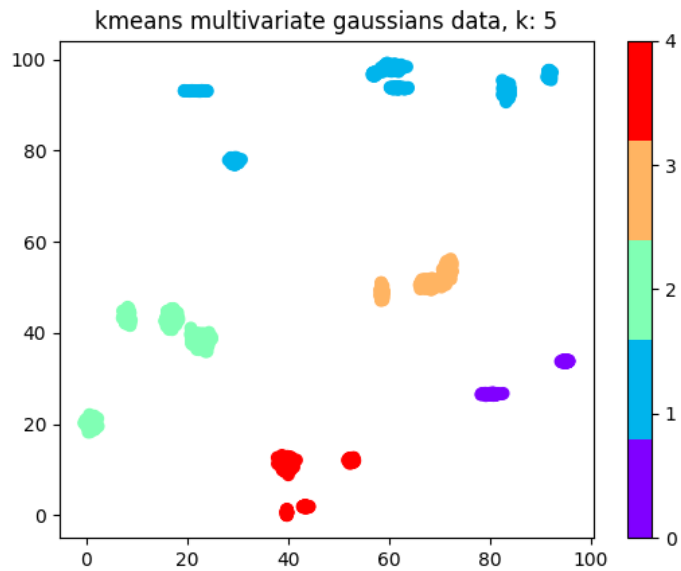
על מנת למצוא את הk המתאים ביותר לדאטא השתמשתי בשיטת "המרפק".
הדאטא הסינטטני שעבדתי עליו היה דאטא שייצרתי לבד ־ יצרתי k גאוסיאנים, עבור כל אחד דגמתי תוחלת ושונות.
על מנת שאוכל להציג את התוצאות לאחר הרצת הקלאסטר עשיתי את זה בדו מימד
על מנת להציג כמה טוב כל אחד מהקלאסטרים השתמשנו באלגוריתם מהכיתה כלומר: $\sum_x ||x-\mu_k||^2$, כאשר μ_k הוא מרכז הקלאסטר שבחרנו לx.
כמו שדיברנו בכיתה אנחנו נצפה לראות גרף יורד ממש, כאשר באיזשהו שלב הירידה תתמתן ובקושי נקבל שיפור, כלומר כל הקלאסטרים יחסית הדוקים ואנחנו ממשיכים לפצל בתוך הקלאסטר ולא בין קלאסטרים שונים.
נשים לב שיכול להיות שהירידה תיהיה לפני שנגיע למספר הקלאסטרים, המצופה כיוון שיכול להיות שהקלאסטרים יחסית קרובים...
תוצאות:



כפי שציפנו קיבלנו גרף יורד ממש כאשר בסביבות 10 או 12 הוא ממש מתייצב, אבל הירידה הגדולה קוראת עד $k = 5$, נסתכל על כמה אופציות כדי לבחור את הקלאסטר הכי טוב

נשים לב שיש לנו כמה נקודות "שונות" שהן לא יורדות ממש, הסבר אחד לזה יכול להיות מכיוון שיש לנו פה אלמנט רנדומי של אתחול המשתנים...

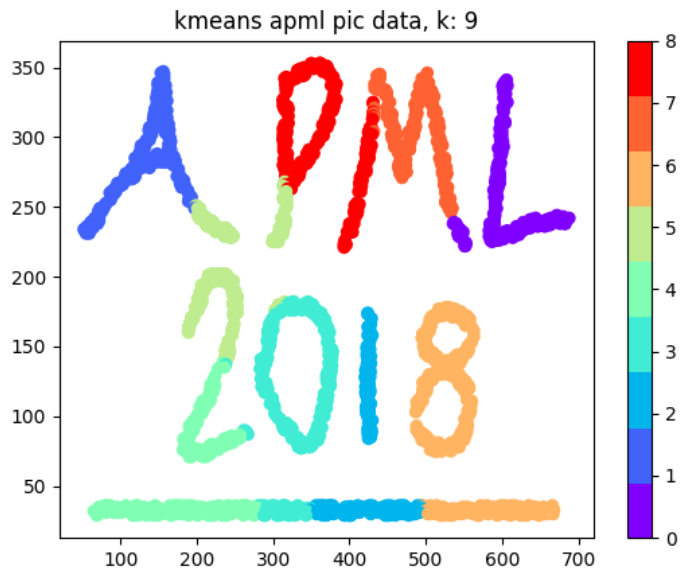




אנחנו רואים שהירידה בין 15 ל5 כן משתלמת במקרה הזה, והצבעים של הקלאסטרים די תואמים את מה שאנחנו כבני אדם היינו עושים...

אלגוריתם Kmeans - על תמונת APML

על מנת להמשיך שהאלגוריתם אכן עובד, אציג את האלגוריתם גם על התמונה של APML



היינו רוצים לחלק לקלאסטרים אחרים, אבל עם היכרות עם השיטה של kmeans ידענו שהוא לא יצליח, כיוון שהמרחקים בין האותיות לפעמים קטנים ולעומת זאת האותיות גדולות, כלומר סביר שלא נראה חלוקה ממש של האותיות.

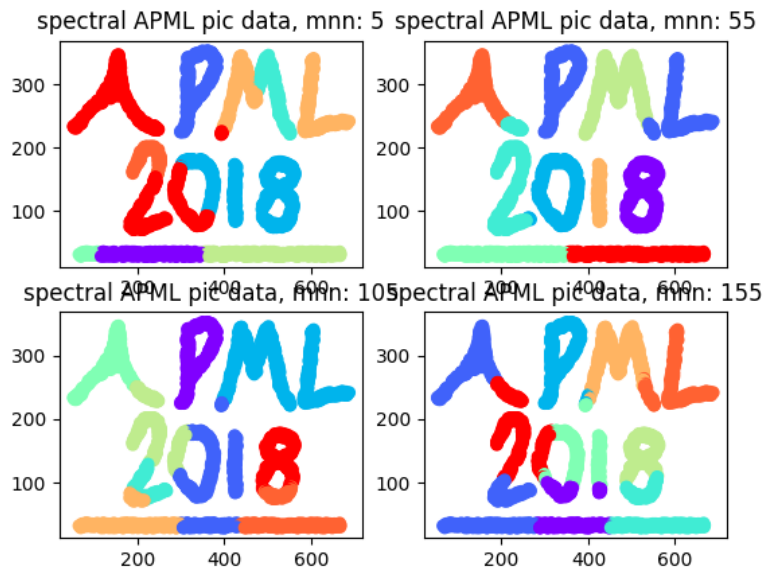
נקווה שהספקטרלי יצליח יותר :)

פירוק ספקטרלי

כמו כן מיממשתי את אלגוריתם הפירוק הספקטרלי, עם 2 סוגי של נורמליזציה (כמתבקש בתרגיל):

שכנים

לא דיברנו בכיתה על אלגוריתם סגור לבחור את m לכן עשיתי "ניסוי וטעיה" ציירתי את הקלאסטרים עם כמה m שונים בטווח יחסית רחב וקיבלתי:



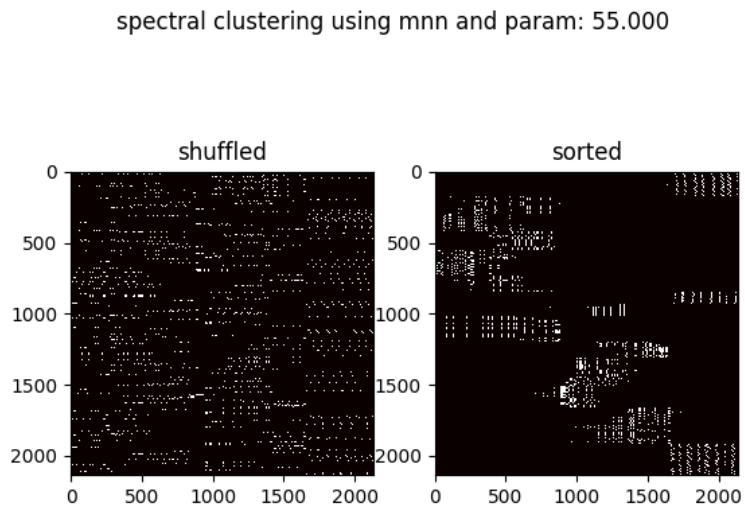
כמובן שזה לא המון סוגים שונים של m וזה כנראה לא תיהיה בחירה אופטימלית, אבל היא תעזור לנו להבין את הכיוון...

ע"פ הציורים האלה בחרתי $m = 55$, כיוון שה"קיבוץ" (clustering) לא מתבצע ב$\mathbb{R}^2$ אלא במימד של הע"ע קשה להסביר למה פרמטר אחד טוב מאחר, אנחנו כן רואים בתוצאות ש:

- עבור $m = 5$ יש נקודות אדומות במקומות שלא היינו מצפים (למשל בתחית הM), כמו כן הצלע הימנית של M מופרדת משאר האות והאות הבאה
- עבור $m = 105$ השבירה של 2 מוזרה, גם האות 0 לא נכנסה לקלסטר בודד
- עבור $m = 155$, שוב $2, 0, 8$ התפצלו למספר קלאסטרים

אבל זה די באוויר, ומישהו אחר היה יכול לבחור פרמטר שונה... אין לנו פה איזשהו מדד מרחק או פונקציית מחיר שיכולה להכריע כל וויכוח בנושא.

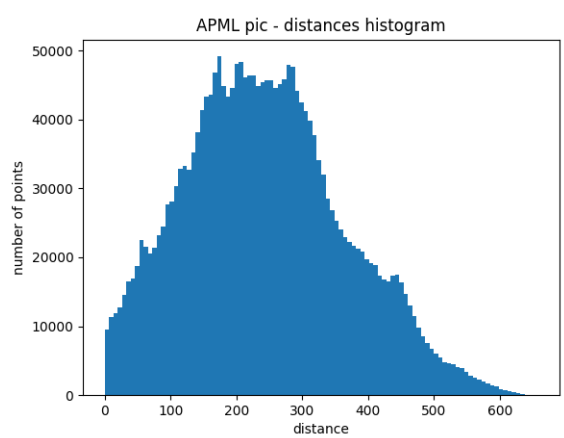
בנוסף ציירתי מפת חום שמתארת את מדד הקרבה (כמו שהתבקשנו בסעיף: 2.4.1)



זה בשחור לבן מכיוון שאנחנו מסתכלים על השכנים הקרובים ביותר - כלומר כל ערך במטריצה יהיה $1,0$ רואים די יפה שבצד שמאל הכל מבולגן, ובצד ימין יש יחסית סדר ואנחנו רואים כמה בלוקים שמציינים על עמודות ושורות שנמצאות באותו קלאסטר והם אכן עם ערכים דומים.

קרנל גאוסיאני

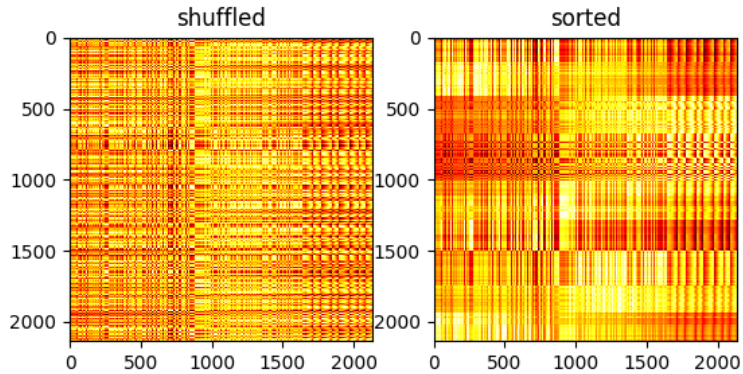
על מנת לבחור את σ השתמשתי בשיטה שראינו בכיתה, ציירתי את היסטוגרמת המרחקים:



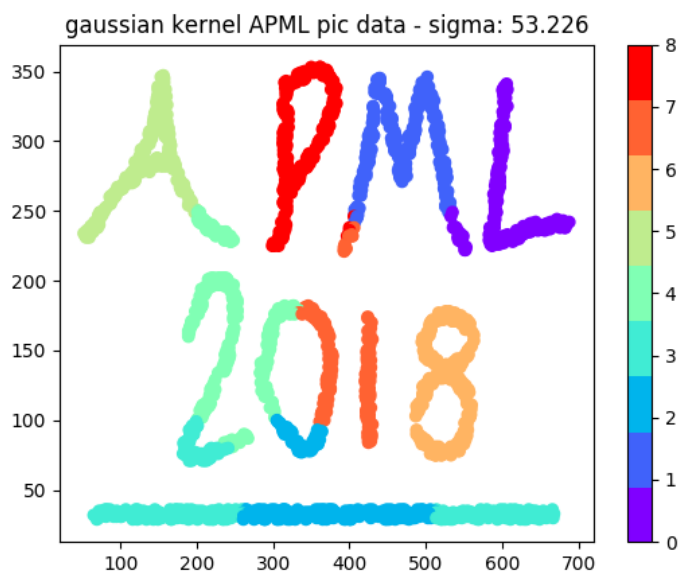
כיוון שהדאטא מתפלג יחסית גאוסיאני וכמו שראינו בכיתה, בחרתי את σ לפי אחוזון $.5$

גם כאן ציירתי את מפת החום, הפעם הערכים הם לא $0,1$ בגלל סוג הפעולה שעשינו (הפעלנו קרנל גאוסיאני שרק ממצע את הערכים):

spectral clustering using gaussian_kernel and param: 53.226



והחלוקה לקלאסטרים שנקבל במקרה הזה היא:



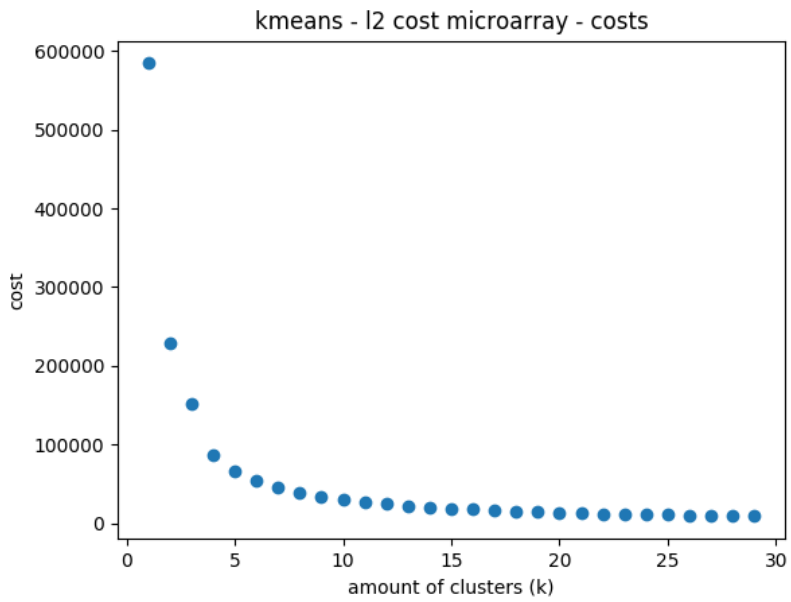
דאטא ביולגי

על מנת להציג את הדאטא הביולגי ציירתי קורלציות בין כל התנאים, הסיבה לכך היא שעשינו את הקלאסטרים במימד הגבוהה ולכן נרצה להשוות גם במימד הגבוה.

אז לקחנו את הדאטא המקורי, ומיינו אותו לפי הקלאסטרים, כלומר $|C_0|$ השורות הראשונות זה קלאסטר 0, $|C_1|$ השורות הבאות מקלאסטר 1 וכן האלה...

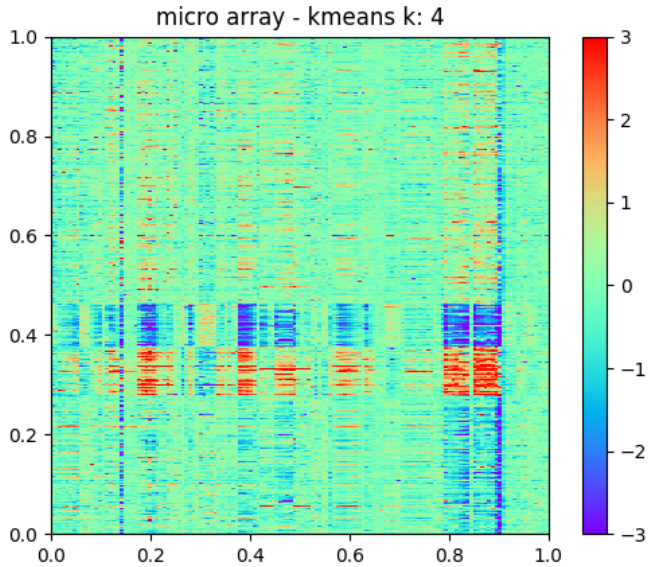
אלגוריתם kmeans

שוב השתמשנו בשיטת המרפק על מנת לבחור את הטווח של k:



אנחנו רואים צניחה משמעותית סביב ב$4-5$ קלאסטרים הראשונים, ואחרי זה כל קלאסטר משפר אותנו בפחות (עדיין בהמון אבל באופן יחסי בפחות).

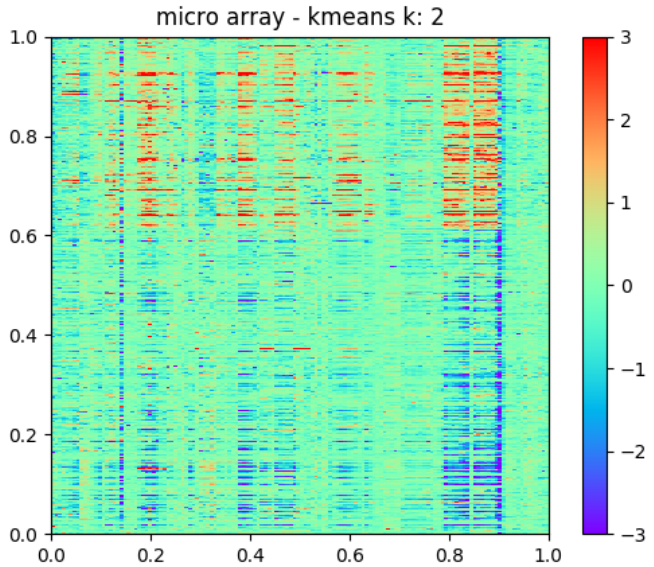
על מנת לבדוק מה הקלאסטר הכי טוב עשיתי מספר נסיונות עם k שונים, הנה חלק מהתוצאות:

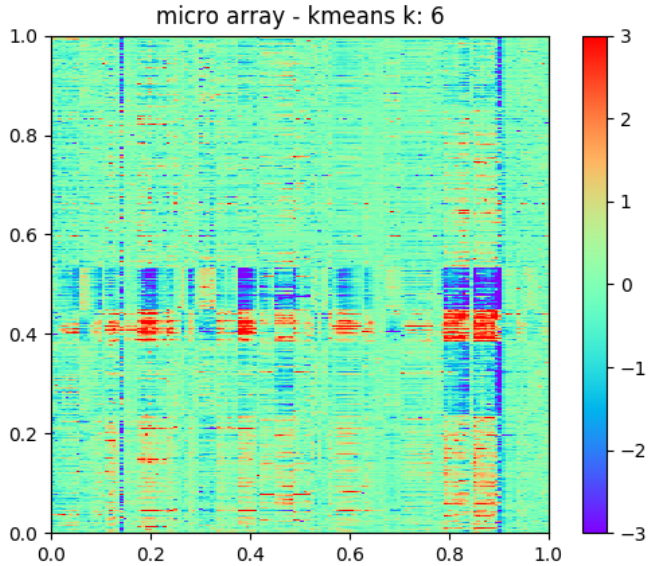


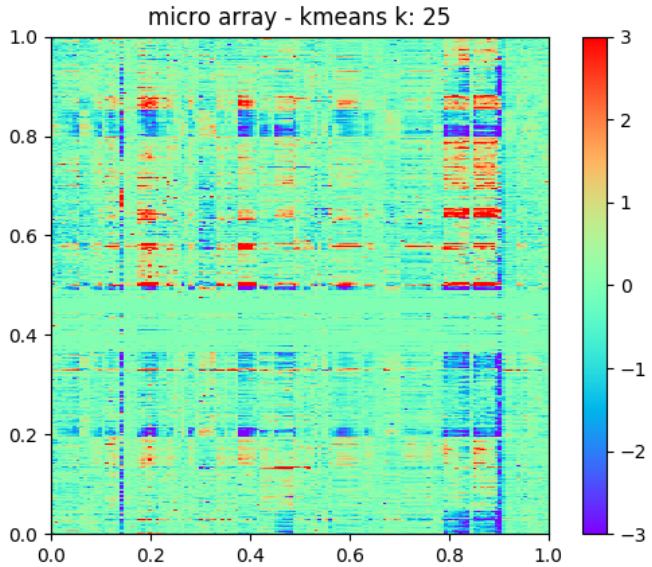
נתמקד ב4 וננסה להבין מה קורה בו (בעזרת זה גם נסביר למה הוא הכי טוב) אנחנו רואים 2 קלאסטרים די מובהקים, אחד בין $0.3 - 0.5$ בצבעים סגולים לאורך כמעט כל התנאים הגנים מתבטאים בצורה דומה וחלשה (הגנים השייכים לקלאסטר הזה), מיד אחריו (כלפי מטה) אנחנו רואים עוד קלאסטר שבו הגנים דווקא מתבטאים בצורה חזקה.

חות משני אלה, אנחנו רואים עוד שני קלאסטרים, אחד מעל הסגול שהוא הקלאסטר הגדול ובו רוב הערכים נמצאים יחסית באמצע ועוד קלאסטר מתחת לאדום גם בו כמעט הכל באמצע חוץ מעמודה די רחבה באזור $0.8 - 0.9$ (בציר x) שם הערכים יחסית נמוכים יותר.

כאשר אנחנו משתמשים בk שונים אנחנו רואים שהם מתחלקים עדיין ל4 הקלאסטרים האלה רק האחרים טיפה יותר מופרדים אבל אין לזה ממש בסיס על סמך מה שאנחנו רואים כאן... לכן אם נרצה הכללה טובה יותר רצוי מספר קטן יותר של קלאסטרים

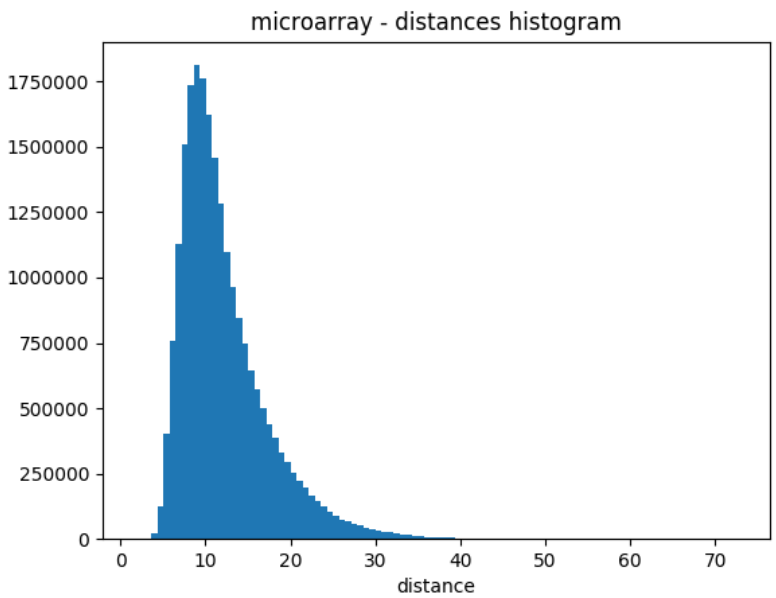

micro array - kmeans k: 2
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
3
2
1
0
−1
−2
−3


micro array - kmeans k: 6
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
3
2
1
0
−1
−2
−3



אלגוריתם spectral

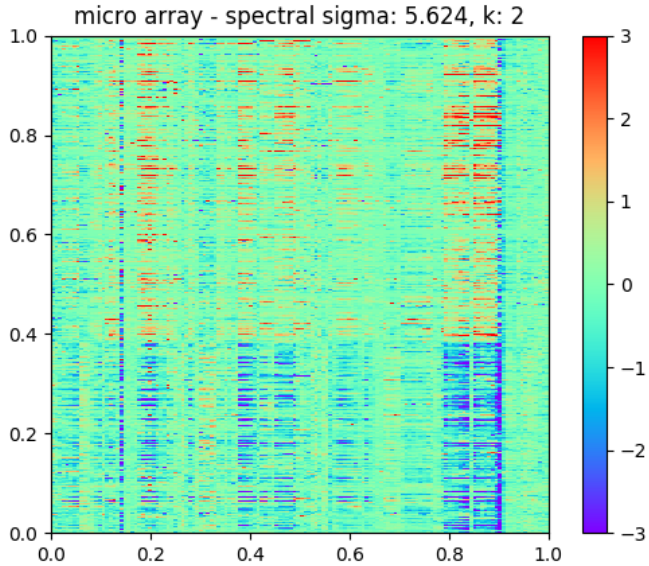
ראשית לבחירה σ, כאמור ציירתי את היסטוגרמת המרחקים:

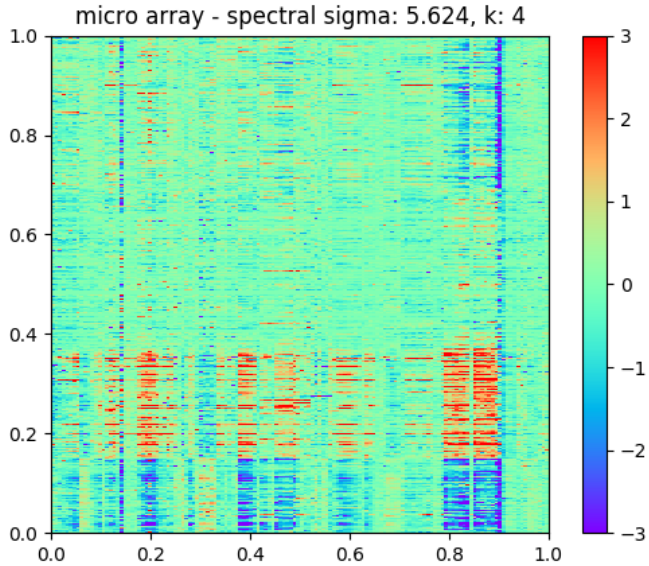


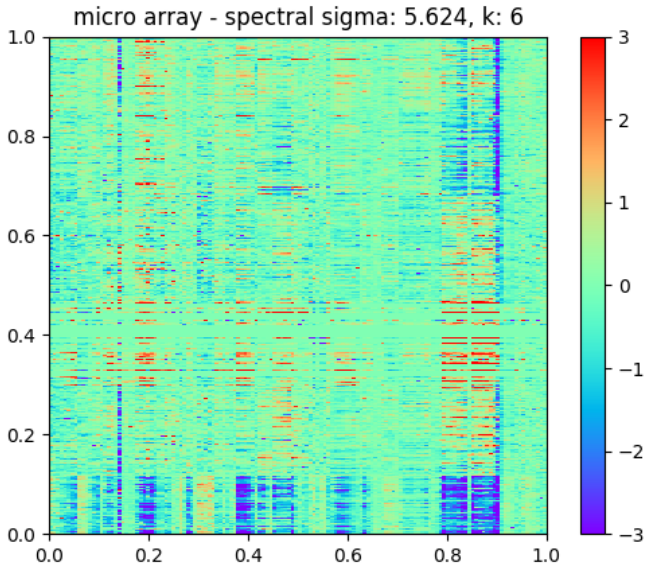
הפעם ההתפלגות היא לא גאוסיאנית, אלא עם זנב ימני ארוך.

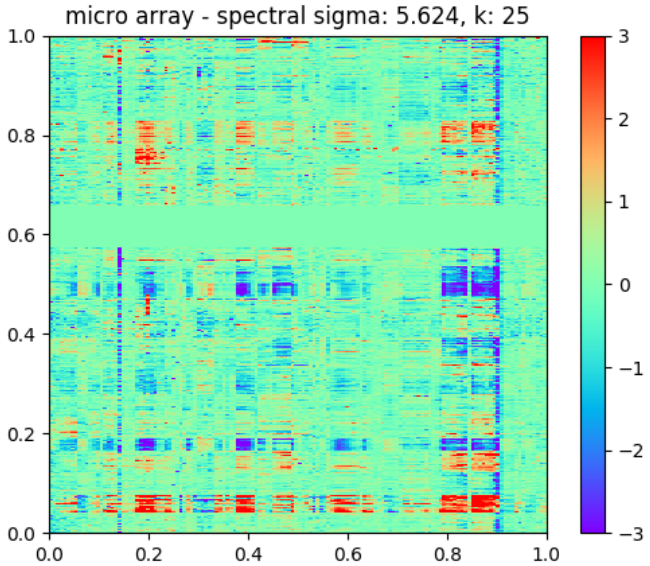
כלומר מצד שמאל של הפיק העלייה היא חדה יותר, ולכן הפעם בחרתי את σ להיות האחוזון ה2 ולא ה5 כפי שהומלץ בכיתה.

מבחינת k, עשיתי מספר ניסויים שאפשר לראות כאן:





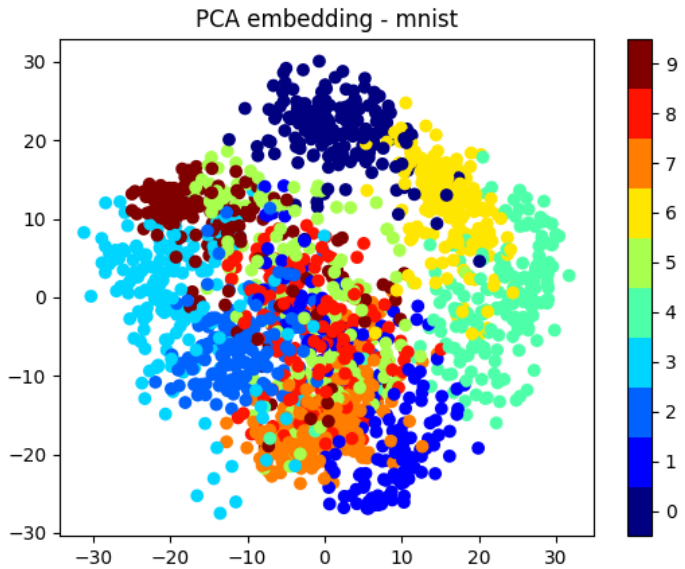


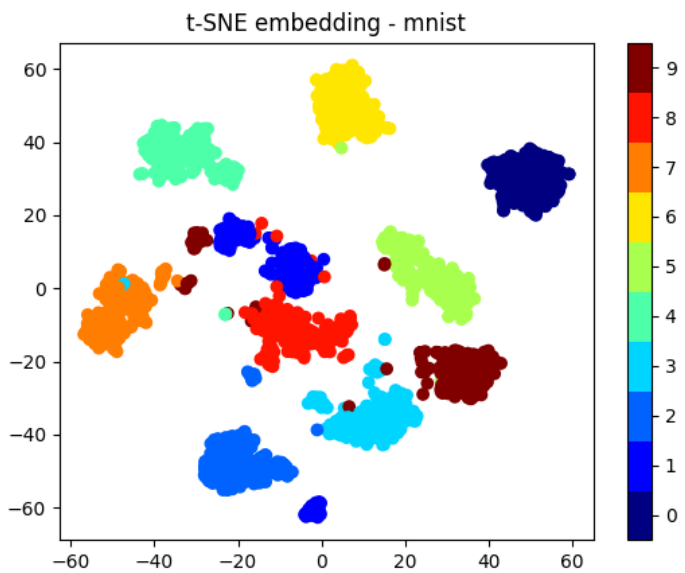


תוצאות די דומות לkmeans, אין לי מה להוסיף :)

הורדת מימד - t-SNE

כנדרש בתרגיל הרצתי את הדאטא של MNIST עם t-SNE ועם PCA והנה התוצאות:





ד

כפי שניתן לראות בPCA קשה לראות הפרדה בין הספרות השונות...

לעומת זאת עם t-SNE ישנה הפרדה בין הספרות, כלומר הורדת המימד שמרה על התכונות של הספרות השונות כמו שהיינו מצפים, זה ההבדל בין האלגוריתמים הם פותרים בעיות אופטימיזציה שונות t-SNE - שמירה על המרחקים במימד הנמוך, PCA הסברה של כמה שיותר מהשונות של הדאטא...